Supplementary Note

When the two genes for alcohol dehydrogenase (Adh 1 and Adh 2) are aligned, and the nucleotides at the silent sites of the two fold redundant codon systems compared, one discovers that the fraction of sites where the nucleotides are the same (f_2) is 0.848. If there were no codon bias, and if the time since the two genes diverged were long relative to the reciprocal of the rate constant with which these silent sites suffer transition substitutions (that is, if enough time has passed to equilibrate the nucleotides at those sites), then $f_2 \approx 0.50$, with a variance reflecting the number of sites (fewer sites implies larger relative variance).

The f_2 metric was introduced (Benner, 1998) to provide a more homogeneous molecular clock than is provided by the dS metric of Yang, a metric that was used by Lynch and Conery to analyze paralogs in the yeast genome. The metric was proposed to be preferable because it counted only transitions, not the mixture of transversions and transitions that determines the dS metric. Its formal mathematical simplicity also enables it to be used more conveniently in reconstructing ancestral transition rate constants.

The f_2 metric is an example of a molecular clock. In yeast, it is widely believed that molecular clocks should not work well. There are a variety of reasons for these. First, gene conversion occurs in yeast. This would tend to keep a pair of paralogs (in the same genome, therefore able to "talk to" each other) from diverging as fast as a pair of orthologs (in different genomes, therefore free to diverge without conversion preserving similarities).

Thus, one does not expect to be able to calibrate a clock for dating paralogs from an analysis of orthologs in yeast, or in any other taxa collection where gene conversion is frequent. Further, codon bias is strong in some genes in yeast. Therefore, even at equilibrium, f_2 is expected to be greater than 0.50, and perhaps as high as 0.60 in some strongly biased genes (Benner, 2003).

Therefore, when the f_2 clock was applied to paralogs of the yeast genome, it was not expected to identify anything significant. Therefore, it was surprising that an analysis of the yeast genome using the f_2 metric discovered a cluster of paralogs in yeast with $0.80 < f_2 < 0.86$ (Benner et al. 2002). Embedded in the middle of this cluster was the Adh 1/Adh 2 paralog pair.

A histogram showing this cluster was published in *Science* in 2002 (Benner et al. 2002). The cluster of paralogs where $0.80 < f_2 < 0.86$ was rather cleanly isolated from duplications with lower f_2 values (many of whose silent codons had equilibrated, that is $f_2 \approx 0.55$) and from duplications with higher f_2 values (many with f_2 values near unity).

Note that the histogram in *Science* contained 16 pairs where $0.80 < f_2 < 0.86$, while the Table 2 has only 15. The 16th pair in the *Science* paper was a pair from the mitochondrial genome. One should not mix nuclear and mitochondrial gene pairs, as the rate of mutation in the first is much slower than the second. Therefore, including this pair in the

Science histogram was a mistake. This mistake was corrected in this manuscript.

The surprising result came when it was observed that the gene pairs in $0.80 < f_2 < 0.86$ window were not randomly selected from the genome, and not associated with the block duplications then being advanced as the principal historical event that shaped the yeast genome. Two thirds of the duplications in the $0.80 < f_2 < 0.86$ window (6 out of 9, generating 11 of the 15 pairs) were involved in the production of ethanol from hexose in the make-accumulate-consume strategy displayed by modern yeast.

Even here, the enzymes were not random. Rather, the enzymes displaying a duplication in the $0.80 < f_2 < 0.86$ window were those that were rate determining and product determining in yeast, and/or were up regulated when the strategy was operative. Bioengineers have long known that one cannot increase the flux of glucose to ethanol by increasing the level of expression of triose phosphate isomerase, hexokinase, or aldolase (for example). In the yeast genome, genes for these proteins are not duplicated. Rather, the duplications generating paralogs having $0.80 < f_2 < 0.86$ were for enzymes such the hexose transporters, believed to be rate determining in the massive movement of sugar to ethanol, glyceraldehyde-3-phosphate dehydrogenase, one isozyme of which increases in expression two fold when yeast is grown on glucose (McAlister & Holland, 1985), pyruvate decarboxylase, which is a product determining step, the thiamine transport (which imports a vitamin needed for pyruvate decarboxylase), and the alcohol dehydrogenases themselves.

Likewise, it is clear that the glyceraldehyde phosphate dehydrogenases have different patterns of expression and different coupling to cell oscillations. Three paralogs exists in *Saccharomyces*: TDH1, TDH2, and TDH3. The double deletion mutants in either TDH1 and TDH3 or TDH1 and TDH2 render the yeast incapable of growth on glucose as the sole carbon source. Interestingly, the activity of the isozymes varies widely, with TDH1, TDH2 and TDH3 representing 10-15, 25-30, and 50-60% of the activity in the cell (McAlister and Holland 1985). A phylogenetic analysis of TDH genes from related yeast shows a similar distribution to that of the PDC genes. The relative proportions of the paralogs for the TDH genes in *S. cerevisiae* are the same in glucose or ethanol, but activity is two-fold higher in glucose grown cells.

One cannot dismiss the idea of f_2 as a molecular clock, suggested that the correct analysis must consider all enzymes involved in the fermentation. The logic of such an argument might be that if *all* enzymes in the make-accumulate –consume ethanol strategy were represented by paralog pairs in the yeast genome, and if most of the duplications that generated these pairs did not have f_2 values in the window, then the appearance of a number of them in the f_2 window would not be statistically significant.

In fact, a majority of enzymes in the fermentation pathway are not represented by paralog pairs in the yeast genome. These include:
Hexokinase (no relevant paralogs) II HXKII
Glucose-6-phosphate isomerase (no relevant paralogs, just PGI1)

PFk (no relevant paralogs, just the two subunits of the heterooctameric enzyme)
Aldolase (no relevant paralogs, just FBA1)
Triose phosphate isomerase (no relevant paralogs, just TPI1)
Phosphoglycerate kinase (no relevant paralogs, just PGK1)

Two enzymes directly involved in glycolysis remain that have duplications outside of the $0.8 < f_2 < 0.86$ window. These are enolase and phosphglycerate mutase. Enolase has two paralogs (ENO1 and ENO2) that have an f2 value of 0.946. These are distantly related to a homolog known as ERR1, with silent sites equilibrated. Phosphglycerate mutase has three paralogs, GM1, GM2 and GM3. The silent sites are essentially equilibrated in these, and the number of characters is small.

Thus, two facts are clear:
1. A majority (6 of 8) of the duplications of proteins involved in the make-accumulate-consume strategy in the $0.8 < f_2 < 0.86$ window
2. A majority of the duplications (6 of 9) in the $0.8 < f_2 < 0.86$ window are involved in the make-accumulate-consume strategy.

This might, of course, be a coincidence. It is it not, then this implies that the f_2 value can be used as a clock (with further caveats to be discussed below). Further, it suggests that the yeast genome contains a record of the emergence of the make-accumulate -consume strategy in the form of the near contemporaneous creation, by duplication of the genes that were needed to implement it, as indicated by paralog pairs having $0.80 < f_2 < 0.86$.

These implications are controversial, but only because the community believes, for the reasons outlined above, that clocks cannot work in yeast.

Given the substantive reasons to doubt that a molecular clock would ever work in yeast, we initially viewed this observation as falling into the category of those that are "too good to be true". This is especially so since in molecular evolution, "contemporaneous" is a statement having the large uncertainties. Thus, the f_2 value is calculated by comparing a finite number of silent sites. Therefore, the value is associated with a variance. For example, an f_2 value of 0.83 calculated from 100 characters has a variance of approximately ± 0.02 (recognizing that the variance is, in fact, asymmetrically distributed). Values calculated from fewer characters have larger variances. For this reason, the f_2 values are generally calculated for a paralog pair only if the number of characters is greater than a certain threshold. In the histogram in *Science*, for example, pairs were included only if they had at least 100 characters aligned, and that the PAM distance separating the two paralogs was less than 120.

If the assumptions behind a clock (e.g., gene-invariance of rate of change) are false, the clock will be overdispersed (see Cutler 2000 a,b). Thus, the fact that one observed a functionally significant cluster suggests that the assumptions behind the clock are not very bad, at least for these gene pairs. This too was a surprise.

But when did these duplications all occur? The cluster of paralogs where $0.80 < f_2 < 0.86$

had f_2 values modestly higher than that displayed by *Saccharomyces-Kluyveromyces* ortholog pairs. This implied that if $f_2$could be used as a clock, and if the clock ticked at the same rate for paralogs within the *Saccharomyces* lineage as it ticked in the *Saccharomyces-Kluyveromyces* orthologs, then this cluster of paralogs would have emerged soon after the divergence of the *Saccharomyces* and *Kluyveromyces* genera.

But when did the *Saccharomyces* and *Kluyveromyces* genera diverge. Correlating the genomic record with the geological record is exceptionally difficult. These difficulties arise from uncertainties in the dates given to fossils in the paleontological record, and the difficulty of finding in that record "transitional forms". This is, of course, especially true for organisms that are not easily fossilized (shelled organisms and vertebrates are not in this category; yeasts are).

Nevertheless, a decade ago, Berbee and Taylor did a heroic task of attempting to date fossil divergences, which are used as the basis for some of the speculations in the discussion of this paper. Assuming that the Berbee-Taylor dating scheme for fungi is approximately correct, this implies that the *Saccharomyces-Kluyveromyces* species pair diverged approximately 100 million years ago (only the first digit in this number is likely to be significant). This implies that the paralog pairs clustering in the window where $0.80 < f_2 < 0.86$ would have diverged shortly thereafter, sometime in the late Cretaceous.

This was also an observation that was too good to be true. Fleshy fruits appeared in the fossil record also during the Cretaceous. Fleshy fruits are the only resource that is sufficiently rich in sugar as to make the make-accumulate-consume strategy worth while. The time of appearance of fleshy fruits is poorly constrained, for the same reason that most dates are poorly constrained in the fossil record. Nevertheless, it is clear that fleshy fruits did not arise before 125 Ma, or after 65 Ma.

Many might be dissatisfied, as are we, with the imprecision in the dates, and the broad uncertainty in time. Even given this broad range in time, however, the range does not include the time when yeast was domesticated by humans. Therefore, even an imprecise clock suffering from these uncertainties in the fossil record, should permit us to distinguish between an event occurring during the Cretaceous and an event occurring in the Pleistocene.

It is possible, of course, that paralog pairs diverge more slowly, because gene conversion is possible between paralogs, than orthologs, between which gene conversion is not possible (Gao & Innan, 2004) . This consideration does not, however, offer a mechanism by which the paralog pairs clustering in the window where $0.80 < f_2 < 0.86$ would have diverged more recently. If there is substantial gene conversion, the consequence would make it appear as if the paralog pairs diverge more recently than they actually did, by comparison with the f_2 clock applied to orthologs pairs. Thus, while gene conversion can make an event that actually occurred in the Pleistocene appear to have occurred in the Cretaceous, it cannot make an event that actually occurred in Cretaceous the appear to have occurred in the Pleistocene. In short, if f_2 can be used as the basis for a clock, we are forced to conclude that the divergence of Adh 1 and Adh 2, as well as all of the paralog

pairs clustering in the $0.80 < f_2 < 0.86$ window did not occur as a consequence of human domestication.

Further, while only a limited number of sequences are available from *Saccharomyces* species other than *cerevisiae*, some genes can be found in the database that permit us to say that the duplications where $0.80 < f_2 < 0.86$ have occurred before the divergence of *bayanus* (for example) from *cerevisiae*, and after the divergence of *Saccharomyces* from (for example) *Kluyveromyces*. Several of these are noted in Table 2. This is true for the pyruvate decarboxylases, for example.

This collection of data was sufficient to make a purely bioinformatics argument that the yeast genome contained the record of the emergence of the make-accumulate-consume strategy near the time when fleshy fruits emerged. Nevertheless, we treated this as a *hypothesis*. The experiment reported in this paper was designed to test this hypothesis, by asking whether the behavior of an ancient protein that was the last common ancestor of Adh1 and Adh2 behaved as if it lived before the make-accumulate-consume strategy arose. It did, and that is the principal conclusion of the paper. Obviously, a series of paleomolecular reconstructions can be done for the other enzymes that duplicated to generate pairs where $0.80 < f_2 < 0.86$.

This conclusion does not require the f_2 metric to be a clock, Conversely, as a single example, It does weakly support the f_2 clock.

The relation between these duplication and block duplications

Further controversy has emerged because the observation that the yeast genome, in its paralogs, might contain a record of a functional adaptation in the past has run afoul of the emerging believe that block duplications dominate the structure of the modern yeast genome. Block duplications should generate, at two points in the genome, syntenic strings of paralogs, that is, two strings of paralogs ordered in the same way in two places.

It requires some degree of bioinformatics sophistication to detect these, as the strains are interspersed with genes that do not have paralogs in the duplicated string. The common explanation for this is that following block duplication, most of the paralog pairs were redundant, and one member of the pair was lost from one string or the other. A summary of the controversy can be found in Kellis et al. (2004).

The block duplication hypothesis comes in several forms. In its weakest form, it holds simply that blocks of genes can duplicate, and that this has happened repeatedly throughout the evolution of the yeast genome. The stronger version of this hypothesis is that many of the block duplications occurred at the same time, generating an event known as the whole genome duplication (or WGD for short). The strongest version of the hypothesis is that *all* duplications occurred at the same time, that is, that whenever one sees a pair of paralogs in the yeast genome, they were generated at the same time as the generation of all other parallel pairs in these genome; that is there was exactly one whole genome duplication.

It should be emphasized that the existence of block duplications, the weakest form of a hypothesis, is well supported. It should also be noted that the existence of a single whole genome duplication in yeast is not from proven, the portions of double synteny between the genome from *Kluyveromyces* and the *Saccharomyces cerevisiae* genome (Kellis et al. 2004) can be interpreted as evidence those blocks were created at the same time.

This notwithstanding, it appears to be universally accepted that duplications could have occurred before or after any WGD, and that duplications can occur for single open reading frames. In principle, it should be possible over time to sequence enough genomes from various yeasts that have diverged at various times from the lineage leading to *Saccharomyces cerevisiae* to identify branches in the ancestral genome history where each block duplication occurred.

What is remarkable about the pairs of paralogs having $0.80 < f_2 < 0.86$ is that if they *are* involved in the make-accumulate-consume strategy, then they are *not* associated with a block duplication. Conversely, if a paralog pair *is* associated with a block duplication, then the paralog pair is *not* associated with the make-accumulate-consume strategy.

We can, of course, use block duplications to test the f_2 metric. Unfortunately, the test is not very interesting. If the Wolfe blocks are examined using the f_2-TREx metric, a pattern is observed. For most duplicated pairs, the silent sites are nearly equilibrated (that is, $0.4 < f_2 < 0.6$) for all of the paralog pairs within all of the blocks. There is, of course, an occasional f_2 value that is more than one standard deviation outside of the mean, as expected given the variance.

There is a consistent exception: ribosomal proteins. Below is one set of data, from Block 8, which has three ribosomal proteins, and which illustrates this quite clearly:

YBR169C	SSE2	536512	heat shock protein	YPL106C	SSE1	1151221	cochaperone	0.56
YBR172C	SMY2	536518	cytoskeleton organization	YPL105C		1151222	function unknown	0.58
YBR177C	EHT1	536527	function unknown	YPL095C		1151231	function unknown	0.54
YBR181C	RPS6B	536534	ribosome protein	YPL090C	RPS6A	1151236	ribosome protein	0.97
YBR182C	SMP1	536538	Trans factor MADS box F	YPL089C	RLM1	1151237	transcription factor	nr
YBR183W		536540	Phytoceramidase	YPL087W	YDC1	1151239	dihydroceramidase	0.53
YBR189W	RPS9B	536552	ribosome protein	YPL081W	RPS9A	2347168	ribosome protein	0.88
YBR191W	RPL21A	536555	ribosome protein	YPL079W	RPL21B	1147614	ribosome protein	0.88
YBR197C		536567	function unknown	YPL077C		1147616	function unknown	0.45
YBR199W	KTR4	536571	mannosyltransferase	YPL053C	KTR6	1079689	mannosylphosphate transferase	0.60
YBR205W	KTR3	536583	mannosyltransferase	YPL053C	KTR6	1079689	mannosylphosphate transferase	0.56

The program did not report an f_2 for the YBR182C/YPL089C pair because it contained too few characters; it is also equilibrated.

It is possible, of course, that the higher f_2 values for ribosomal proteins is coincidental. Given the consistency of this observation, this explanation seems unlikely. It appears more likely that ribosomal proteins have higher amounts of gene conversion and/or codon bias than other classes of proteins, or that divergence at silent sites is less likely to be

neutral. We can easily imagine conjectures to explain why this might be so. But these considerations suggest that the f_2 metric fails to correctly date the divergence of ribosomal proteins, relative to other proteins.

We might also ask whether synteny is a 100% reliable indicator of contemporary divergence. With 90% of the genes following the putative WGD having disappeared, with duplications occurring not in blocks, and with duplications occurring throughout the history of the genome, there is more than ample opportunity for pieces of synteny to converge.

We can ask about the distribution of f_2 values in the paralogs that form the blocks. Of the 348 paralog pairs in the Wolfe blocks (including those where the number of characters is fewer than 100), the distribution is shown by the following Table:

Window	f_2 values
0.300-0.399	4
0.400-0.499	52
0.500-0.599	171
0.600-0.699	60
0.700-0.799	16
0.800-0.899	25
0.900-0.999	19
1.0	1

The population from 0.3 to 0.8 is consistent with a typically overdispersed clock. The excess of pairs having $0.8 < f_2 < 1.0$ is consistent with a second mode. The fact that these are dominated by ribosomal proteins cannot be an accident. The occasional non-ribosomal protein in this segment (for example, a pair of homocitrate synthases in the blocks has an $f_2 = 0.89$) is consistent with a rare appearance in a block by random chance.

Two of the duplications in the $0.80 < f_2 < 0.86$ windowthat are not associated with the fermentation pathway are indeed associated with blocks. They are small blocks. The f_2 values display the same pattern, where most are equilibrated silent sites, but one is not. The f_2 data are shown below:

seqid1	seqid2	f2	n2		
gi|1370464	gi|854447	0.545	66	YPL224C	YMR177W
gi|1370472	gi|854450	0.583	36	YPL228W	YMR180C
gi|1370474	gi|854451	0.654	26	YPL229W	YMR181C
gi|1370480	gi|854453	0.514	107	YPL232W	YMR183C
gi|1370495	gi|854456	0.810	311	YPL240C	YMR186W

seqid1	seqid2	f2	n2		
gi|1323227	gi|849162	0.600	60	YGR136W	YPR154W

| gi\|1323230 | gi\|849164 | 0.860 | 171 | YGR138C YPR156C
| gi\|1323236 | gi\|849165 | 0.564 | 101 | YGR141W YPR157W
| gi\|1323238 | gi\|849166 | 0.514 | 37 | YGR142W YPR158W
| gi\|1323238 | gi\|849166 | 0.514 | 37 | YGR143W YPR159W

When considering these data, one needs to recognize that the number of characters (n2) used to calculate the f_2 value for many of these pairs is low. Thus, an f_2 value calculated from 37 silent sites has a large variance.

These results may be used to criticize the f_2 clock, or to criticize the notion that if genes form a syntenic block, that they must have duplicated at the same time. In our view, these are simply examples where in two (of hundreds) of pairs, the f_2 value is more than a few standard deviations away from the mean. This is expected from the variance, and is a feature that limits the usefulness of any clock based on discrete characters, including f_2. It appears that the duplications in the $0.80 < f_2 < 0.86$ window that are not involved in the make-accumulate-consume strategy, and are associated with block duplications, lie more than two standard deviations outside of the variance.

But as these are not related to any of the pairs that are involved in the make-accumulate-consume ethanol strategy, resolving this discrepancy is not relevant to this paper.

In short, the f_2 data suggest that duplications in the $0.80 < f_2 < 0.86$ window occurred after most of the Wolfe blocks were created. The f_2 values show consistent equilibration, with the expected variance, which generates an occasional f_2 value more than one SD outside of the mean. The f_2 clock consistently fails with ribosomal proteins, if the blocks are correctly assigned, and we might generate some reasonable suggestions as to why this is.

References for the supplemental material

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F. & Cullin, C. A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **21**, 3329-3330 (1993).

Benner, S. A., Trabesinger-Ruef, N., Schreiber, D. R. Post-genomic science. Converting primary structure into physiological function. *Adv. Enzyme Regul.* **38**, 155-180 (1998).

Benner, S. A., Caraco, M. D., Thomson, J. M., Gaucher, E. A. Planetary biology. Paleontological, geological, and molecular histories of life. *Science* **293**, 864-868 (2002).

Benner, S. A. Interpretive proteomics. Finding biological meaning in genome and proteome databases. *Adv. Enzyme Regul.* **43**, 271-359 (2003).

Berbee, M. L., Taylor, J W. Dating the evolutionary radiations of the true fungi. Canadian Journal of Botany **71**: 1114-1127 (1993).

Bozzi, A., Saliola, M., Falcone, C., Bossa, F. & Martini, F. Structural and biochemical studies of alcohol dehydrogenase isozymes from *Kluyveromyces lactis*. *Biochim. Biophys. Acta* **1339**, 133-142 (1997).

Cutler, D. J. Estimating divergence times in the presence of an overdispersed molecular clock. Mol Biol Evol **17**:1647-1660 (2000a)

Cutler, D. J. 2000b. Understanding the overdispersed molecular clock. Genetics **154**:1403-1417 (2000b).

Gao, L. & Innan, H. Very low gene duplication rate in the yeast genome. *Science* **306**, 1367-1370 (2004)

Guthrie, C. & Fink G. R. Guide to yeast genetics and molecular biology. *Methods in Enzymology* **194**, eds. Abelson, J. N & Simon, M. I. Academic Press, Inc. Harcourt Brace Jovanovich, Publishers. New York (1991).

Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**, 617-624 (2004).

Larroy C, F., M. R. Fernandez, E. Gonzalez, X. Pares, & Biosca, J. A. Characterization of the *Saccharomyces cerevisiae* YMR318C (ADH6) gene product as a broad specificity NADPH-dependent alcohol dehydrogenase: relevance in aldehyde reduction. *Biochem. J.* **361**, 163-172 (2002a).

Larroy, C. F., X. Pares, & Biosca, J. A. Characterization of a *Saccharomyces cerevisiae* NADP(H)-dependent alcohol dehydrogenase (ADHVII), a member of the cinnamyl alcohol dehydrogenase family. *Eur. J. Biochem.* **269**, 5738-5745 (2002b)

McAlister, L. & Holland, M. J. Differential expression of the three yeast glyceraldehydes-3-phosphate dehydrogenase genes. *J. Biol. Chem.* **260**, 15019-15027.

Thomson, J. M. *Interpretive Proteomics: Experimental Paleogenetics as a Tool to Analyze Function and Discover Pathways in Yeast.* Ph.D. Dissertation, University of Florida, Gainesville, FL U.S.A. (2002).

van Der Wel, H., Morris, H. R., Panico, M., Paxton, T., North, S. J., Dell, A., Thomson, J. M. & West, C. M. A non-Golgi alpha 1,2-fucosyltransferase that modifies Skp1 in the cytoplasm of Dictyostelium. *J. Biol. Chem.* **276**, 33952-33963 (2001).